Defining a verb taxonomy by a decision tree

Eric Laporte

Defining a verb taxonomy by a decision tree

Éric Laporte
Universidade Federal do Espírito Santo
Université Paris-Est, LIGM (UMR 8049)

Abstract

How useful can be decision trees for the construction, update and use of taxonomies of verbs? They are supposed to make the definition of classes more accurate. They are still little used in practice: devising and handling them involves formal logic and Boolean reasoning, which do not belong to a background in humanities. However, a better organization of interdisciplinarity in linguistic research might actually overcome such reluctance. We exemplify how the decision tree of the Lexicon-Grammar of French verbs led us to revise this taxonomy. We resolved an internal contradiction in eight classes: with some locative verbs, a content clause can occupy the theme argument, and we made the formalization of this feature consistent with the definitions of classes.

Introduction

This article discusses how useful can be a decision tree for the construction, update and use of a taxonomy of verbs.[1]

Verb taxonomies are produced for two main objectives: as a contribution to the description of lexical features of verbs, and as a part of syntactic and semantic databases for language processing. In both contexts, decision trees are supposed to make the definition of classes more accurate. They are still little used in practice: devising and handling them involves formal logic and Boolean reasoning, which do not belong to a background in humanities. However, a better organization of interdisciplinarity in linguistic research might actually overcome such reluctance. We exemplify how the decision tree of the Lexicon-Grammar of French verbs (Tolone, 2012) led us to revise this taxonomy, resolving an internal contradiction in eight classes: with some locative verbs, the theme argument can contain a content clause, and we made the formalization of this feature consistent with the definitions of classes.

Section 1 introduces the purpose of verb taxonomies, and what decision trees are supposed to provide for them. Section 2 surveys related work. Section 3 analyses the reluctance of specialists with respect to decision trees. Section 4 reports the work on locative works achieved with the aid of a decision tree.

1. Purpose of verb taxonomies and decision trees

Verb taxonomies are produced for two main purposes, which belong to linguistics and to language processing.

In linguistics, taxonomies contribute to describing lexical features of verbs. Classes are defined on the basis of common syntactic constructions or semantic features. For instance, dative verbs in French like *donner* ‘give’ or *prêter* ‘lend’ can be defined as transitive verbs with a third argument introduced by a preposition, in general *à* ‘to’, and by a few other syntactic and semantic features: clitic pronominalization of the prepositional argument, meaning of exchange between two of the arguments... Defining the class is an opportunity to

[1] This article is dedicated, as a testimony of sincere friendship, to Yoichiro Tsuruga, a life-long connoisseur of French verbs.

study in depth these features. During this process in the case of dative verbs, Leclère (2007), for example, decided to include *de* 'from' among the prepositions introducing the third argument, incorporating into the class verbs as *obtenir* 'obtain'.

In addition, description is simpler when you have classes. For example, describing the verb *emprunter* 'borrow' is easier once a list of features of dative verbs is ready.

In language processing, taxonomies can be used as syntactic and semantic databases. For instance, the Lexicon-Grammar taxonomy of French verbs was recently used as a database of lexical information with the FRMG parser (Tolone, Sagot, 2011).

In both contexts, a decision tree is supposed to make the definition of classes more accurate. Fig. 1 shows a sample of a decision tree associated with a taxonomy of French verbs. In botany, mycology and zoology, decision trees are routinely used to identify species on the basis of their observable features. In the same way, when a new verb is assigned a class, using a decision tree is a way to comply with the definition of existing classes and to ensure that the new entry is placed in the right class.

- N_1 destination place of N_0 ..... 35LD
- (N_1 source place of N_0) or (N_1 static place of N_0)
  - ♦ N_1 *V de* N_0 ..... 34L0
  - ♦ not (N_1 *V de* N_0)
    - N_1 source place of N_0 ..... 35LS
    - N_1 static place of N_0 ..... 35ST
- not (N_1 place of N_0) 35LR

Fig. 1. Excerpt of the decision tree of the Lexicon-Grammar of French verbs (Tolone, 2012). This part deals with a collection of two-argument verbs with a prepositional object. N_0 stands for the argument which is subject of the canonical construction and N_1 for its prepositional object. Sequences in italics stand for syntactic constructions

In particular, what makes decision trees particularly useful is the fact that the classes in a taxonomy should be mutually exclusive. As Gross puts it (1978, 80-81),

> the classifications made by traditional grammarians (...) consist of disjoint classes, i.e. of equivalence classes in the sense of algebra. (...) One could also think of dividing a set into non-disjoint classes, i.e. into subsets whose intersections are not empty. It is clear, however, that this second procedure would be impracticable even for the smallest classes. Whereas it is easy to grasp a partition of elements into disjoint classes (see Fig. 2), it is practically impossible to get a direct perception of the same set, if the classes are divided in a more complex way (see Fig. 3). (...) The perceptual reasons for the difficulty in understanding this latter system of classes are rather evident. Therefore, all important classifications which have been constructed so far are based on disjoint classes. Notice that one way to represent graphically the disjunction of classes is a classificatory tree where the characteristics determine branching.

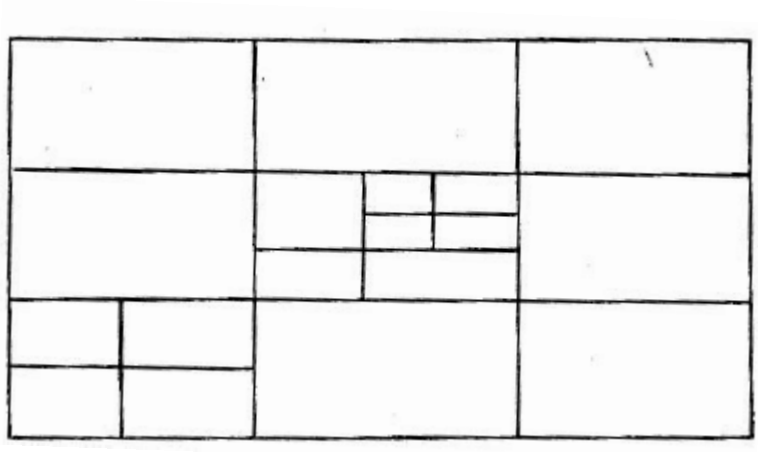

Fig. 2. Division of a set into 18 disjoint subsets. A hierarchy may be established on the basis of the different sizes of the rectangles

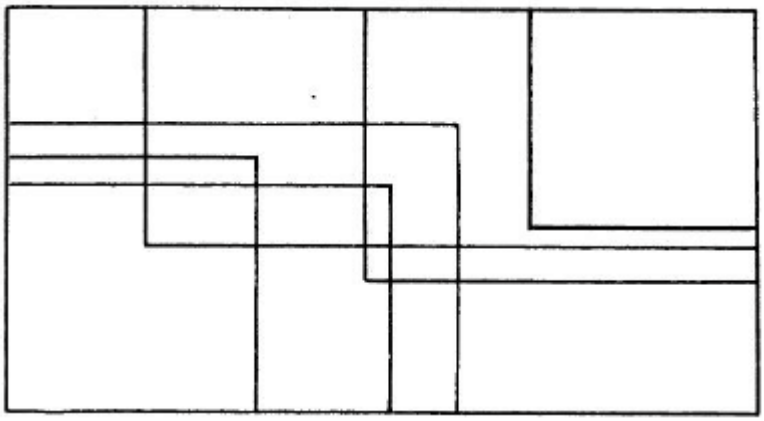

Fig. 3. Division of a set into non-disjoint classes where, for example, each subclass can be considered as a rectangle having two sides on the sides of the largest rectangle. However, other interpretations are equally possible

Gross's requirement is not contested. All major existing verb taxonomies are composed of mutually exclusive classes. This is confirmed by available listings of classes. The assignment of a given verb, with a given interpretation, to distinct classes would unanimously be considered as an anomaly or an error.

The fact that classes should be disjoint leads to devising decision trees. Boons *et al.* (1976:28) explain this connection on the example of a set of 'conversion' verbs (our translation):

> Each class is essentially defined by a feature, which is naturally called the 'definitional feature' of the class. In the particular case of [some set of 'conversion' verbs], definitional features do not cut out disjoint classes in the set of verbs in question. In other words, a single verb use can cumulate the definitional features of several classes. Such a verb use is categorized in only one of these classes, which is therefore considered as a priority over others. (...) the order of priority is shown in the following diagram.

The diagram mentioned does not specify all the relevant priorities between definitional features, but if it did, it would translate straightforwardly into a decision tree.

At each node of a decision tree, all options are mutually exclusive, i.e. each option must exclude all its siblings. For example, in Fig. 1, 'N_1 source place of N_0' and 'N_1 static place of N_0' are offered as options: the decision tree assumes that the N_1 argument may not be

simultaneously source place and static place of N_0. More specifically, no lexical entry may allow N_1 to be interpreted both as source place and static place of N_0. If this happened, for example for *partir* 'go away', the verb would be described with separate lexical entries, one with N_1 as source place, as in *La tache part de la chemise* 'The stain washes out of the shirt', and the other with N_1 as static place, as in *Le chemin part de la gare* 'The path begins at the station'.

Many nodes in a decision tree are binary, and offer two options that are the negation of each other, as 'N_1 *V de* N_0' and 'not (N_1 *V de* N_0)' in Fig. 1. This ensures that the options are mutually exclusive. When there are more than 2 options, as in the uppermost node of Fig. 1, checking that they are mutually exclusive requires a little more logical reasoning. In any case, logical negation plays an important part, in connection with the fact that definitions are both a matter of inclusion and exclusion.

In terms of information content, a decision tree is equivalent to the definitions of all the classes it mentions, with the only difference that it avoids repeating a criterion used in several definitions, e.g. 'not (N_1 *V de* N_0)', which is definitional for both classes 35LD and 35ST. Therefore, the tree can serve as a reference in the same way as the definitions themselves:

- when describing constructions and identifying classes, it order to assign a class to a new entry;

- when assessing or revising an existing description;

- when using descriptions, in order to deduce information from the membership of an entry in a class.

In other words, decision trees seem to be helpful for practically all human tasks involved by the management of verb taxonomies.

2. Related work

The good-sense observations above give the impression that decision trees are a convenient tool in any verb taxonomy, to the point that using them is a quality requirement. Are they used in current practice? Let us review major verb taxonomies by order of creation date.

In the Lexicon-Grammar (Gross 1994), each class is defined by criteria, but these definitions are not always precise enough that they arbitrate in case of doubt between two classes. Authors and users have a mental image of each class and assign verbs to them by comparison (Leclère 2007). Take, for example, class 4 of Gross (1975): it is defined by the basic construction

(1) Class 4 : *Qu P V* N_1

where *Qu P* stands for a content clause and N_1 for a direct object; class 6 is defined by

(2) Class 6 : N_0 *V Qu P*

where N_0 stands for the subject. Since slots N_0 and N_1 may be filled by content clauses, these definitions do not specify whether verbs with the basic construction *Qu P V Qu P* should be assigned to class 4 or 6. As a matter of fact, *ridiculiser* 'ridicule' in

(3) *Que vous soyez émotif ridiculise que vous ayez l'ambition de jouer au poker*
'Your being emotional ridicules your ambition to play poker'

is assigned to class 4, and *mériter* 'deserve' in

(4) *Que vous ayez réussi mérite que nous fassions une fête*
'Your success deserves a party'

to class 6. The definitions tolerate an overlap that actual descriptive practice excludes clearly. Only recently, in collaboration with Christian Leclère and Elsa Tolone about the Lexicon-Grammar of French verbs (Tolone 2012), we formalized the mental images of classes into a decision tree, through a study of cases where the delimitation between classes was not explicitly documented in literature. This tree is available on the LIGM's website.[i]

The case of VerbNet, which incorporates Levin's (1993) taxonomy, is similar. The verbs with the same or similar alternation behaviour are assigned to the same class; however, definitions mention only acceptable constructions and not inacceptable ones. Compare, for example, the definitions of classes 11.1 and 11.2: in addition to shared constructions, 11.1 is characterized by one additional construction:

(5) *NP V NP PP.destination PP.initial_location*

and 11.2 by four others. This does not logically rule out the possibility that a verb would belong to both classes with the same interpretation. As a matter of fact, *scoot* is listed in 11.2, but accepts (5) in

(6) *The balloon scooted to the far right from the left*

We did not find published studies about such overlaps between VerbNet classes.

Comlex (Macleod *et al.* 1995) classifies syntactic constructions rather than lexical entries. For example, *ask* is assigned at least 20 constructions to account for *John asked a favour, John asked me a favour, John asked for advice, John asked me for advice, John asked Tom over, John asked for there to be a proctor at the exam, John asked me to drive him, The beggar asked me to eat a bite, John asked that they pay attention, John asked to eat, I asked if I should go, John asked when the blizzard would come, They asked him what to do, They asked about his leaving early, We asked about them eating the food, I asked about whether he would come, They asked of her that she remain, He asked of everybody if he could have a piece or not, I asked him about what to do, I asked him about their offering him more money, I asked him about no one having been there*… without delimiting how many lexical entries are to be formalized for this verb. This makes the problem quite different, since one of the reasons why one can hesitate between two classes for a verb is that two of its constructions (with the same sense) point to different classes. The Comlex guidelines (Wolff *et al*. 1998) do not define the syntactic constructions with a decision tree, but generally provide specific criteria when a doubt between two constructions is possible.

In FrameNet (Johnson, Fillmore 2000), class definitions rely mainly on semantic intuition and therefore are relatively fuzzy: a verb sense is assigned to the frame which descriptors feel to be closest. This procedure practically evacuates hesitations between classes.

The generative framework is a special case: it discourages the definition of word taxonomies. Features of words are listed as in Fig. 4, but are not ranked according to any notion of relevance or hierarchy. 'On the basis of such lexical representations, it is impossible to discover any organization of the lexicon (Gross 1978, 83).'

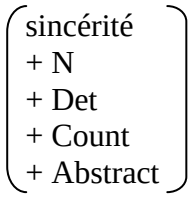

Fig. 4. List of word features in the generativist style.

Summing up, actual classes are disjoint in any taxonomy, but, paradoxically, none of the major existing verb taxonomies did systematically attempt to work out or refine definitions so as to rule out overlaps between classes; in particular, none used decision trees until recently.

3. Reluctance with respect to decision trees

In order to explain why projects of verb taxonomies are so reluctant about decision trees, let us focus on the authors of the taxonomies and of the would-be trees. The linguistic criteria involved in their daily work are elaborate linguistic notions, e.g. the distinction between arguments and adjuncts. These researchers are required to have linguistic theoretical knowledge, and to be trained to sophisticated practical skills such as corpus study or acceptability judgment, often both. Such specialists, in general, have a complete academic background in linguistics.

Decision trees involve the same notions and skills, of course, but also formal logic. Let us examine how much formal logic is involved on a concrete example of two definitions allowing for an overlap between the corresponding classes, namely (1)-(2). The 'contested' verbs are those that accept content clauses in both the N_0 and N_1 arguments. Suppose we wish to give priority to the N_0 *V Qu P* construction of class 6 over the *Qu P V* N_1 construction of class 4, i.e., we allocate all of the contested verbs to class 6. We would have to reformulate the definition of class 4 by inserting a negative condition which forbids the N_1 argument to take the form of a content clause:

(7) Class 4: (*Qu P V* N_1) and not (*Qu P V Qu P*)

This option could also be expressed in the form of a decision tree:

- N_0 *V Qu P* 6
- not (N_0 *V Qu P*)
  Qu P V N_1 4
  not (*Qu P V* N_1) ..... (other classes)

Both the definitions and the decision tree ensure logically that the two classes do not overlap. In the case of the definitions, the logical proof of this is: if a verb belonged to classes 4 and 6, its basic construction would be described by both *Qu P V* N_1 and N_0 *V Qu P*, implying *Qu P V Qu P*, which is explicitly ruled out by (7). Note that both the definitions and the decision tree involve explicit logical negation.

Now, allocating the *Qu P V Qu P* overlap to class 6 was a simplistic option without linguistic motivation, just for the sake of illustration. In fact, in the Lexicon-Grammar listings, the verbs in the overlap are divided into the two classes: *mériter* 'deserve' is assigned to class 6, but *ridiculiser* 'ridicule' to class 4. If we examine how the listings discriminate these verbs, we can notice a correlation with infinitival constructions:

- If N_0 can be an infinitive clause controlled by N_1, the verb belongs to 4. Example: *Tenir ces propos a ridiculisé Paul* 'Saying that ridiculed Paul'.

- If N_1 can be an infinitive clause controlled by N_0, the verb belongs to 6. Example: *Votre succès mérite d'être fêté* 'Your success deserves a commemoration'.

In addition, these two features about infinitival constructions are widespread among lexical entries of, respectively, classes 4 and 6. We can use them for a second try at adjusting the definitions of the classes, with a better linguistic motivation. In order to express this adjustment in precise definitions, we introduce Lexicon-Grammar notations for infinitive clauses. 'N_0 can be an infinitive clause controlled by N_1' formalizes as N_0 =: V^1-*inf W*, where '=:' means 'can take the form of', *V-inf* stands for a verb in the infinitive, the superscript

number specifies co-reference with another argument in the sentence, and *W* symbolizes a possible sequence of complements.

In an attempt to adopt simple definitions on these bases, we might try these:

(8) Class 4: (*Qu P V* N_1) and (N_0 =: V^1-*inf W*)
(9) Class 6: (N_0 *V Qu P*) and (N_1 =: *de* V^0-*inf W*)

These two definitions are simple and look symmetrical, but they have several shortcomings.

Firstly, the classes may theoretically still overlap: if we come up with a verb that accepts both infinitival constructions, it might satisfy both (8) and (9). There is no obvious example of such a case, but there might come to be one. As a matter of fact, the logical form of definitions (8) and (9) show no attempt at making the classes disjoint.

Secondly, if a verb accepts none of the infinitival constructions, it is excluded from both classes, as *autoriser* 'allow':

(10) *Paul autorise que Rose parte* 'Paul allows Rose to leave'
(11) **Paul autorise de partir* '?*Paul allows to leave'
**Être malade autorise que Rose parte* '*Being ill allows Rose to leave'

This is because definitions (8) and (9) do not explicitly state that the criteria about infinitival constructions should be taken into account only for verbs with the basic construction *Qu P V Qu P*. We give definitions a logical formal in order to make them precise; this logical formalization requires that no information is left implicit.

Finally, definitions (8) and (9) are too restrictive as compared to the listings of the classes: they restrict classes 4 and 6 to make verbs that accept the corresponding typical infinitival constructions, whereas a few verbs do not, e.g., in class 4, *asphyxier* 'suffocate':

Ce type d'éducation asphyxie les enfants 'This type of education suffocates children'
**Être trop protégés asphyxie les enfants* 'Being overprotected suffocates children'

and, in class 6, *autoriser* 'allow' (10)-(11). Again, we should have formalized that the criteria about infinitival constructions are to be taken into account only for verbs with the basic construction *Qu P V Qu P*.

In order to adjust the two definitions in a more satisfactory way, we need to be more careful with logical aspects. Firstly, we will take into account the infinitival constructions only when we are in the *Qu P V Qu P* overlap. We can thus sketch a decision tree:

- *Qu P V* N_1
  not (N_0 *V Qu P*) ..... 4 pro parte
  N_0 *V Qu P* ..... (overlap between 4 and 6, to be continued)
- not (*Qu P V* N_1)
  N_0 *V Qu P* ..... 6 pro parte
  not (*Qu P V* N_1) ..... (other classes)

Note that we had to resort to logical negation in line 2.

Secondly, when in the *Qu P V Qu P* overlap, we must decide about verbs that admit none of the infinitival constructions, like *autoriser* 'allow', or both, if any such verb exists. These verbs look syntactically and semantically closer to class 6 than to class 4, and most are listed in class 6. Therefore, we can expand line 3 of our tree as in Fig. 5. In the latter ♥ line of Fig. 5, we do not need to discriminate whether the 'N_1 =: *de* V^0-*inf W*' construction is observed or not. If it is, we are in the case of *mériter* 'deserve', and assign the verb to class 6. If it is not, then the verb admits none of the infinitival constructions, and we also assign it to

class 6. Since the decision is the same in both cases, the distinction would needlessly complexify the tree. Note how we used logical reasoning to formulate the branches for verbs that admit none or both of the infinitival constructions.

- *Qu P V* N_1
  - ♦ not (N_0 *V Qu P*) ..... 4 pro parte
  - ♦ N_0 *V Qu P*
    - ♥ N_0 =: V^1*-inf W*
      - not (N_1 =: *de* V^0*-inf W*) ..... 4 pro parte
      - N_1 =: *de* V^0*-inf W* ..... 6 pro parte
    - ♥ not (N_0 =: V^1*-inf W*) ..... 6 pro parte
- not (*Qu P V* N_1)
  - N_0 *V Qu P* ..... 6 pro parte
  - not (N_0 *V Qu P*) ..... (other classes)

Fig. 5. Sketch of a decision tree delimiting classes 4 and 6.

Let us make a final adjustment to the pair of definitions. Some verbs behave syntactically like *mériter* ‘deserve’, except that they admit another infinitival construction:

Paul démontre avoir atteint le pôle	‘Paul proves he reached the pole’
**Paul démontre d’atteindre le pôle*	‘*Paul proves of reaching the pole’

or an interrogative clause:

Paul élucide si Rose a atteint le pôle	‘Paul elucidates whether Rose reached the pole’
**Paul élucide d’atteindre le pôle*	‘*Paul elucidates of reaching the pole’

The corresponding features, formalized as N_1 =: *Aux*^{0}*-inf V W* and N_1 =: *si P ou si P*, should be used in the same way as N_1 =: *de* V^0*-inf W*, which leads us to Fig. 6. Nearly all nodes in this decision tree are binary, and offer two options that are the negation of each other.

- *Qu P V* N_1
  - ♦ not (N_0 *V Qu P*) ..... 4 pro parte
  - ♦ N_0 *V Qu P*
    - ♥ N_0 =: V^1*-inf W*
      - N_1 =: *de* V^0*-inf W* ..... 6 pro parte
      - N_1 =: *Aux*^{0}*-inf V W* ..... 6 pro parte
      - N_1 =: *si P ou si P* ..... 6 pro parte
      - none of the preceding ..... 4 pro parte
    - ♥ not (N_0 =: V^1*-inf W*) ..... 6 pro parte
- not (*Qu P V* N_1)
  - N_0 *V Qu P* ..... 6 pro parte
  - not (N_0 *V Qu P*) ..... (other classes)

Fig. 6. Decision tree delimiting classes 4 and 6 (Tolone, 2012).

A decision tree can be converted into equivalent definitions, i.e. without changing the assignment of verbs to classes. In general, this conversion involves combining various criteria into a Boolean formula. For example, the definition of class 4 becomes:

(12) *Qu P V* N_1 and (not (N_0 *V Qu P*) or (N_0 =: V^1*-inf W* and not (N_1 =: *de* V^0*-inf W* or N_1 =: *Aux*^{0}*-inf V W* or N_1 =: *si P ou si P*)))

Definition (12) contains logical negations as often as the decision tree. In addition, it has parentheses embedded on three levels, and we cannot dispense with them: this would make the formula ambiguous. The decision tree was more readable, after all.

As this example shows, devising and handling decision trees involves formal logic and Boolean reasoning. Few linguists are familiar with such knowledge or trained to such skills, which do not belong to a background in humanities. Specialists of verb taxonomy do not usually feel like undertaking the construction of decision trees: this is not surprising. When computer scientists need syntactic lexicons for their applications, they also avoid to engage in 'tedious and time-consuming' (as many of them say) tasks of linguistic description.

In addition to the fact formal logic is little attractive for linguists, some of the objectives that could lead them to build decision trees are alien to their preoccupations. Recall that decision trees, or precise definitions of classes, facilitate the assessment and revision of existing taxonomies. But this type of task is unfrequent in the current community of linguists, and is practically always performed by the own authors of the taxonomies. We did not find a single research paper aiming at assessing the quality of one of the syntactic dictionaries cited in Section 2, and not signed by one of the authors of the dictionary. Papers proposing revisions or extensions of existing taxonomies are only a little more frequent (e.g. Danlos, Sagot, 2009; Nakamura, 2009). This suggests that linguists, as a scientific community, behave as if it did not fully make sense for an author to assess or revise the results of an other's research. In biology or physics, as soon as important experimental results are published, they are checked, and sometimes completed, by independent laboratories. Such practice in linguistics would increase consciousness about quality assessment and result improvement.

Another potential motivation to create or handle decision trees is related to the use of taxonomies: with precise definitions of classes, users can deduce information from the membership of an entry in a class. Unfortunately, this does not seem to motivate linguists either: projects using language resources are generally language-processing projects, managed by computer scientists independently of the authors of the resources. More familiarity between the two fields would benefit to language processing, but such changes take time.

Globally, the reasons that fuel reluctance to decision trees appear to be 'bad' reasons, in that a better organization of interdisciplinarity in linguistic research might actually overcome them.

4. Using a decision tree to revise a taxonomy

We have mentioned above why decision trees are useful for the management of a taxonomy of verbs. We can exemplify one of these uses: our decision tree of the Lexicon-Grammar of French verbs (Tolone, 2012) led us to revise this taxonomy.

The tree evidenced an internal contradiction in eight classes of locative verbs: 38L0, 38LH, 38LHD, 38LHS, 38LHR, 38LD, 38LS, 38LR of Guillet, Leclère (1992). The definition of all these classes explicitly states that none of the arguments can be occupied by sentential clauses:

Paul a jeté la balle dans la rivière 'Paul threw the ball into the river'
**Paul a jeté qu'il est satisfait dans la rivière*
'*Paul threw that he was satisfied into the river'

However, in the tables describing these classes, the N_1 *=: Qu P* feature encoded the possibility for a content clause to occupy the direct object of the basic construction, and a few verbs were marked as having this feature, for example:

Paul a affiché un message sur la porte 'Paul posted a message on the door'
Paul a affiché sur la porte que nous avons changé de salle
'Paul posted on the door that we changed rooms'

This encoding contradicted the definition of the classes. We undertook to resolve the contradiction. We took care to avoid three potential damaging consequences on the taxonomy:

- losing valuable information,

- weakening the homogeneity of classes,

- or, for lexically ambiguous verbs, impairing the relevance of the distinctions between lexical entries.

4.1. Four typical examples

The presence of the N_1 *=: Qu P* feature in these classes is explained in Guillet & Leclère (1992:418): it recorded in fact the existence of another sense of the same verb, described in another lexical entry, which admits a content clause in the theme argument, N_1, and therefore belongs to another class: 'Since verbs governing content clauses are encoded elsewhere, [the N_1 *=: Qu P* feature] is a column making references to other entries'. For example, the Lexicon-Grammar distinguishes, among other senses of the verb *enregistrer* 'record', those illustrated by the following sentences:

(13) *Le greffier a enregistré la plainte sur son registre*
'The clerk recorded the complaint on his file'
(14) *Paul a enregistré que c'est lundi* 'Paul took in that today is Monday'

The sense illustrated by (13) is listed in class 38LD, whereas (14), an informal use, is described by an entry in class 6 (Guillet, Leclère, 1992:161, footnote). The 38LD entry was marked as having the N_1 *=: Qu P* feature. This indication was a loose reference to the entry for the informal use in class 6. Earlier versions of the Lexicon-Grammar contained other features meant as loose references to other entries, e.g. the *V concret* feature in class 4 of Gross (1975). As the taxonomy was gradually completed, such features became redundant and were removed from the tables.

For some verbs, the other entry was to be created. This is a second kind of situation. Consider for example the following sentences with the verb *mettre* 'put':

(15) *Paul a mis son lit dans un coin* 'Paul put his bed into a corner'
(16) *Paul a mis dans ce paragraphe qu'il est Français*
'Paul put in this paragraph that he is French'

(15) is described in an entry of class 38LD which was marked as having the N_1 *=: Qu P* feature, suggesting a use like (16). However, this metaphorical use was not encoded in any other class. The solution was to create an entry for it in class 10. This was consistent with the content of this class: other lexicalized metaphors were already encoded there, e.g. metaphors transferring the meaning of a verb from space to text, like *glisser* 'slip', *insérer* 'insert' or (15)-(16), or from space to a more or less Freudian view of mind, as in:

Les ouvriers enfouissent des déchets sous la mer (class 38LD)
'The workers bury waste under the sea'
Paul a enfoui dans sa mémoire qu'Ida était partie (class 10)
'Paul buried in his memory that Ida had left'

Other verbs corresponded to a third kind of situation, because N_1 *=: Qu P* was difficult to interpret in the same way as above. The entry *interpoler* of class 38LD is illustrated by

(17) *Des copistes ont interpolé des gloses dans le texte*
'Copyists interpolated glosses in the text'

It had the N_1 *=: Qu P* feature, but no other entry of this verb was encoded in a class of verbs governing content clauses. We know a single use of *interpoler* with a content clause as a direct object, and we can illustrate it by

(18) *Un copiste a interpolé dans le texte que ceci s'est produit en 1083*
'A copyist interpolated in the text that this happened in 1083'

In contrast with the (15)-(16) pair, the verb has exactly the same sense in (17) and (18): the constructions with the nominal direct object and with the content clause should be described within the same entry. In the existing entry, the N_1 *=: Qu P* indication did not really refer to another sense. Apparently, it played the usual part of features in the LG: record a genuine syntactic property of the entry. However, this property reassigns the entry to another class than 38LD. The fact that the authors of the taxonomy had left it in 38LD shows that they felt it was better described there. The definitions of the classes seem to contradict their linguistic intuition about the class where the entry best fits. We will go back to this matter in Section 4.3.

Finally, in a fourth type of situation, some of the data encoded under the N_1 *=: Qu P* feature may be considered as errors. For example, the entry *exhaler* 'exhale' of class 38L0, illustrated by

(19) *La muqueuse exhale de la vapeur d'eau* 'The mucous membrane exhales steam',

had the N_1 *=: Qu P* feature. For us, *exhaler* has no lexicalized use with a content clause as a direct object; the closest to this that we can fabricate is something like:

(20) *Le corps de Paul exhale qu'il brûle de désir*
'Paul's body exudes he is burning with desire'

We judge (20) as a creative metaphor, not as a lexicalized one as (16). As such, it is not to be described in the LG.

Such cases may be explained by material mistakes, or by differences between authors as regards acceptability judgment, especially when it boils down to a distinction between creative vs. lexicalized metaphors.

4.2. Analysis

We studied individually the lexical entries with the N_1 *=: Qu P* feature in the eight relevant classes. Most entries corresponded to one of the 4 cases analysed in the previous section. In cases of type (13)-(14) and (19)-(20), no action was required; in cases of type (15)-(16), a new entry was created in a class of verbs governing content clauses; in cases of type (17)-(18), the entry was transferred to such a class.

The distinction between types (15)-(16) and (17)-(18) was the most difficult. It depends on whether the construction with a content clause, (16) or (18), and the nominal construction, (15) or (16), are to be described by distinct lexical entries or by a single one. When the two constructions have different meanings, two entries must be distinguished, as in the following case:

Paul a posté sa lettre dans cette boite 'Paul posted his letter in this mailbox'

(21) *Paul a posté sur le site qu'il a aimé le film*
'Paul posted on the site that he liked the movie'

(21), an English borrowing, is specific to the Web. An entry was created for it in class 10.

In most cases, comparison of meanings was not decisive, and as it relies on intuition, it does not lead to reproducibility in case of doubt; thus, we often resorted to more formal criteria.

In a few cases, differences in argument structure led us to distinguish two entries. For example, compare the following uses of *couper* 'cut':

(22) *Paul a coupé ce passage de la bande* 'Paul cut this passage from the tape'

(23) ?*Paul a coupé que le président est fou* '?Paul cut that the president is insane'

(22) was described by an entry in class 38LS with the N_1 *=: Qu P* indication. If we accept (23), it has the same meaning, but does not admit the location argument observed in (22):

?**Paul a coupé (de + dans) la bande que le président est fou*
'Paul cut from the tape that the president is insane'

The meaning of the verb is the same, but (23) had to be described in a distinct entry with two arguments, which was created in class 6.

This case was not common. Usually, lexicalized metaphors with a shift from a spatial sense to an abstract sense exhibit no change in argument structure or in the preposition of the location argument (Lamiroy, 1987:51-52). Distributional differences in the theme and location arguments were the most frequent reason which led us to conclude that the construction with a content clause should be described in a distinct entry. For instance, (15) and (16) have the same number of arguments, with the same prepositions:

(15) *Paul a mis son lit dans un coin* 'Paul put his bed into a corner'

(16) *Paul a mis dans ce paragraphe qu'il est Français*
'Paul put in this paragraph that he is French'

However, the distribution of the theme argument (N_1) in the use exemplified by (15) does not always combine with the distribution of the location argument (N_2) in the one exemplified by (16):

**Paul a mis son (lit + stylo) dans ce (paragraphe + discours)*
'*Paul put his (bed + pen) in this (paragraph + speech)'

and vice versa:

?**Paul a mis dans le (placard + coffre) qu'il est Français*
'?*Paul put in the (cupboard + trunk) that he is French'

Lexicalized metaphors from space to text, or from space to mind, often exhibit a joint redefinition of both distributions of the theme and location arguments. In such a situation, the space sense and the metaphorical sense are better described by distinct entries.

Once a case of the (15)-(16) type has been identified, the new entry was assigned to a class. Similarly, for a case of the (17)-(18) type, the existing entry was reassigned to a new class. The class was determined by the decision tree. Most of the entries created or reassigned in this work were listed in class 10, e.g. *mettre* 'put' in (16) and *interpoler* 'interpolate' in (17)-(18). Others were listed in class 6 because the locative argument is not preserved, e.g. *couper* 'cut' in (23).

Some entries listed in the 38LD class are simultaneously examples of the (13)-(14) type and of the (15)-(16) or (17)-(18) type. For example, *afficher* 'post' has, among others, the following uses:

(24) *Paul a affiché un message sur la porte* 'Paul posted a message on the door'
(25) *Paul a affiché sur la porte que nous avons changé de salle*
'Paul posted on the door that we changed rooms'
(26) *Paul affiche qu'il est hétéro* 'Paul flaunts his being straight'

The entry for (24) in 38LD bore the N_1 *=: Qu P* feature. This indication could be interpreted as a loose reference to the already existing entry for (26) in class 6, which has a quite different meaning, or to the construction of (25). The (24)-(26) pair is comparable with (13)-(14). A distributional and transformational analysis of (24) and (25) shows different behaviours according to whether the N_1 argument denotes the material support or the content of the message. If it denotes the support, the N_2 argument excludes computational media:

**Le système a affiché* (*une feuille de papier + un poster*) *dans la fenêtre*
'*The system displayed a (paper sheet + poster) in the window'

If it denotes the content, the N_2 argument can denote computational media:

Le système a affiché (*les résultats + qu'il y a une erreur*) *dans la fenêtre*
'The system displayed (the results + that there is an error) in the window'

Most nouns in N_1 are ambiguous between support and content, and this is reflected in an ambiguity of the sentence if N_2 is ambiguous between a concrete and a computational medium:

Paul a affiché un (billet + message) sur l'écran
'Paul posted a (note + message) on the screen'

Either the message is a paper and the screen a concrete surface, or the message is a text and the screen is an electronic display. These facts suggest distinguishing two lexical entries for (24) and (25). The entry where the N_1 argument denotes a material support was already in 38LD, and the other was created in class 10.

The distinction between types (15)-(16) and (19)-(20) depends on whether the metaphor is lexicalized, as (16), or not, as (20). There is a continuum between these two situations, and the decision depends on introspection. We tried to set the boundary at the same place as in the rest of the LG.

The constructions analysed in this work have in common that they comprise simultaneously a location argument, *sur la porte* 'on the door' in (25), and a theme argument containing a content clause, *que nous avons changé de salle* 'that we changed rooms' in (25). For completeness, we checked two variants of this type of syntactic construction.

The first variant has simultaneously a source location and a destination location. None of the entries with the N_1 *=: Qu P* indication admitted two location arguments:

La chaleur a propagé l'épidémie d'Asie en Europe
'Heat spread the epidemic from Asia to Europe'
**Les réseaux sociaux propagent d'Asie en Europe qu'il y a une pandémie de grippe*
'*Social networks spread from Asia to Europe that there is a flu pandemic'

The other variant has an interrogative clause instead of a *que*-complementized content clause. Guillet & Leclère (1992:161-162) mention such a case with *répertorier* 'list':

Paul a répertorié quels étaient les meilleurs skieurs
'Paul listed which were the best skiers'

They observe that this feature 'makes the entry close to verbs governing content clauses'. According to the LG taxonomy, this feature does even assign the entry to a class of verbs governing content clauses, as in the case of *se demander* 'wonder' in class 6:

Paul se demande s'il pleut	'Paul wonders whether it is raining'
**Paul se demande qu'il (pleut + pleuve)*	'*Paul wonders that it is raining'

For all the entries with the N_1 *=: Qu P* indication, we checked if they admit an interrogative clause in the N_1 argument. Some do, e.g. *mettre* 'put' in (16), but this feature can be encoded in the entry with the *que*-complementized content clause.

4.3. Encoding

Our analyses led us to encode new information in the LG. In this section, we review and discuss the updates decided on this basis.

For cases similar to (16), we created 35 entries in classes 6 and 10. Guillet & Leclère (1992) had detected during their work on locative verbs that these creations were required, and they had stored this information with the N_1 *=: Qu P* feature, but they had delayed the creations. The LG of French verbs was built in a modular way: each author specialized in an object of study and a set of classes, in order to facilitate their task. Most of the new entries are so similar to the existing ones that they are almost duplications (cf. Lamiroy, 1987:51-52): this is the case of (25).

Cases similar to *interpoler* 'interpolate' (17)-(18) are different. Definitions of classes assign this verb to class 10, a class of verbs governing content clauses, but the authors of the LG taxonomy left it in class 38LD, a class of verbs with nominal arguments only, probably feeling it was better described there. This looks like a contradiction between the definitions of the classes and their content: on the one hand, the taxonomy was defined in order to assign each entry to the class where it best fits; on the other hand, its authors felt some entries to be better described in another class than theirs.

What might explain this paradox is that the description of entries does not consist only in assigning them to a class: it also includes a selection of features encoded in the corresponding table. Some important features of *interpoler* were not encoded [ii] in class 10, whereas class 38LD provides them: some are implied by the definition of the class, others are encoded in the table. The only obstacle to placing *interpoler* in 38LD was the content clause. We can thus hypothesize that it was a tempting alternative solution to leave this entry there temporarily. In order to investigate this hypothesis, let us examine a few features provided in class 38LD and not in 10.

By definition, entries in class 38LD exhibit a syntactic and semantic feature related with locative interpretation and which was not encoded in the table of class 10: 'N_2 place of destination of N_1'. This feature is defined and tested by two criteria:

- The N_2 argument answers questions with *où* 'where', e.g.

Où le copiste a-t-il interpolé que ceci s'est produit en 1083 ?
'Where did the copyist interpolate that this happened in 1083?'
– Dans le texte. '– In the text.'

- Support sentences (Guillet, Leclère, 1992:22-25) fabricated from the elementary sentence under study describe situations respectively before and after the process denoted by it:

Que ceci s'est produit en 1083 n'est pas dans le texte
'That this happened in 1083 is not in the text'
Que ceci s'est produit en 1083 est dans le texte
'That this happened in 1083 is in the text'

As a matter of fact, these forms are relevant to the syntactic and semantic description of this entry. In the same way, the definitions of classes 38LH, 38LHD, 38LS, 38LHS, 38LR, 38LHR and 38L0 imply similar features, not encoded in classes of verbs governing content clauses.

In addition, other syntactic features are encoded in these 8 tables and not in class 10: they cumulatively total to 77 features; 34 are related with form and semantics of locative argument, 12 with derived nouns and verbs, 8 with transformations altering the syntactic positions of arguments, etc. Some of these features are relevant to the description of *interpoler* and other entries in the case of the (17)-(18) pair.

Therefore, we have a likely explanation for the paradox of the (17)-(18) pair and similar cases: some features important for them were not encoded in content-clause verb classes, whereas locative verb classes provided them; the only obstacle to placing these entries in locative classes was the content clause. They were left in these classes temporarily.

Another solution is possible now: encoding new features in table 10, and reassigning the entries. By applying the definition of the classes, all are reassigned to class 10.

Therefore, we made a selection of 12 features and encoded them in table 10: 4 about form and semantics of the location argument, 4 about derived nouns and verbs, 3 about transformations altering the syntactic positions of arguments, and the N_1 *=: Nconc* feature, which encodes the possibility of the N_1 argument to denote concrete objects. We transferred 12 entries into class 10. The number of entries created and transferred is displayed in Table 1 for each of the 8 relevant classes of locative verbs.

Entries	38LD	38LS	38LR	38LH	38LHD	38LHS	38LHR	38L0	Total
with the feature	60	44	18	1	10	8	4	27	172
created	16	10	4	1	2	1	0	1	35
transferred	8	3	1	0	0	0	0	0	12

Table 1. Entries with the N_1 *=: Qu P* feature.

After these updates, the N_1 *=: Qu P* feature was deleted from the 8 tables. This feature did not appear elsewhere: in other tables, the acceptability of a content clause in the N_1 argument is indicated by the N_1 *=: Qu Pind* and N_1 *=: Qu Psubj* features, which also specify the mood of the verb in the content clause. Thus, the N_1 *=: Qu P* feature was also deleted from the table of classes and from the documentation on the features.

The updates reported in this article will be available in the next release of the LG of French verbs (version 3.5).[iii]

Conclusion

In spite of the problems posed by the use of formal logic and Boolean reasoning by linguists, decision trees concentrate accurate information about features of lexical entries that specialists of grammar feel as particularly salient. Hence, they deserve attention on the part of linguists committed to formalization or interested in computational applications.

Thus, the decision tree of the Lexicon-Grammar of French verbs (Tolone, 2012) led us to revise the representation of some locative verbs, for which a content clause can occupy the

theme argument. We made the formalization of this feature consistent with the definitions of classes, by creating some entries and transferring a few others into a class of verbs governing content clauses.

References

Boons, Jean-Paul, Alain Guillet, and Christian Leclère. 1976. *La structure des phrases simples en français. Classes de constructions transitives*. LADL Research report no. 6, Université Paris 7.

Danlos, Laurence, and Benoît Sagot. 2009. "Constructions pronominales dans Dicovalence et le lexique-grammaire. Intégration dans le Lefff. " *Lingvisticæ Investigationes* 32(2):293-304. Amsterdam/Philadelphia: Benjamins.

Gross, Maurice. 1975. *Méthodes en syntaxe. Régime des constructions complétives*. Paris: Hermann.

Gross, Maurice. 1978. "Taxonomy in Syntax." *SMIL, Journal of Linguistic Calculus* 3-4: 78-96. Stockholm: Skriptor.

Gross, Maurice. 1994. "The Lexicon-Grammar of a Language. Application to French." In *Encyclopedia of Language and Linguistics,* ed. by K. Brown. London: Pergamon Press.

Guillet, Alain, and Christian Leclère. 1992. *La structure des phrases simples en français. Les constructions transitives locatives*. Genève: Droz.

Johnson, Christopher R., and Charles J. Fillmore. 2000. "The FrameNet Tagset for Frame-Semantic and Syntactic Coding of Predicate-Argument Structure." *1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000), April 29-May 4, 2000, Seattle WA*, 56-62.

Lamiroy, Béatrice. 1987. "Les verbes de mouvement. Emplois figurés et extensions métaphoriques." *Langue française* 76:41-58, Paris : Larousse.

Leclère, Christian. 2007. "Organization of the Lexicon-Grammar of French Verbs." In *Lexicology: Critical Concepts in Linguistics*, ed. by Patrick Hanks, vol. 4: Syntagmatics, 169-187. London/New York: Routledge.

Levin, Beth, 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Macleod, Catherine, Ralph Grishman, and Adam Meyers. 1995. "Tagging as a Means of Refining and Extending Syntactic Classes." *AAAI Spring Symposium on the Lexicon at Stanford University*, 81-85.

Nakamura, Takuya, 2009. "Sur les arguments sémantiques du verbe expliquer et leur réalisation syntaxique. Descriptions du lexique-grammaire" *Lingvisticæ Investigationes* 32(2):187-199. Amsterdam/Philadelphia: Benjamins.

Tolone, Elsa. 2012. "Maintenance du lexique-grammaire. Formules définitoires et arbre de classement. " *TAL* 52(3):153-190.

Tolone, Elsa, and Benoît Sagot. 2011. "Using Lexicon-Grammar tables for French verbs in a large-coverage parser." In *Human Language Technology. Challenges for Computer Science and Linguistics. 4th Language and Technology Conference, LTC 2009, Poznań, Poland, November 6-8, 2009, Revised Selected Papers*, ed. by Zygmunt Vetulani, vol. 6562 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 183-191. Berlin: Springer Verlag.

Wolff, Susanne Rohen, Catherine Macleod, and Adam Meyers. 1998. *COMLEX Word Classes Manual*. Proteus Project, New York University.

[i] http://infolingu.univ-mlv.fr/english/DonneesLinguistiques/Lexiques-Grammaires/Download.html

[ii] Most of the LG of French verbs was built in the 1970s and 1980s. Publication techniques limited the number of features that could be encoded in tables. For example, 36 features were encoded in table 10. Now, spreadsheets allow for viewing and editing tables with hundreds of columns.

[iii] The LG of French verbs is available under the LGPL-LR license at the LIGM's website: http://infolingu.univ-mlv.fr/english/DonneesLinguistiques/Lexiques-Grammaires/Download.html